Protein Science
DOI 10.1002/pro.2163

Smooth statistical torsion angle potential derived from a large conformational database via adaptive kernel density estimation improves the quality of NMR protein structures

Guillermo A. Bermejo,[1] G. Marius Clore,[2] and Charles D. Schwieters[1]*

[1]Division of Computational Bioscience, Center for Information Technology, National Institutes of Health, Bethesda, Maryland 20892-5624

[2]Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, Maryland 20892-0520

***Correspondence to:** Charles D. Schwieters, Division of Computational Bioscience, Center for Information Technology, National Institutes of Health, 12 South Dr., Building 12A, Room 2041, Bethesda, MD 20892-5624. E-mail: charles.schwieters@nih.gov, phone: 301-402-4914, fax: 301-402-2867

Running title: Smooth statistical torsion angle potential

Number of manuscript pages:	38
Number of tables:	2
Number of figures:	4

Received: Aug 01, 2012; Revised: Sep 14, 2012; Accepted: Sep 17, 2012

Abstract: Statistical potentials that embody torsion angle probability densities in databases of high-quality X-ray protein structures supplement the incomplete structural information of experimental nuclear magnetic resonance (NMR) datasets. By biasing the conformational search during the course of structure calculation towards highly populated regions in the database, the resulting protein structures display better validation criteria and accuracy. Here, a new statistical torsion angle potential is developed using adaptive kernel density estimation to extract probability densities from a large database of more than 10^6 quality-filtered amino acid residues. Incorporated into the Xplor-NIH software package, the new implementation clearly outperforms an older potential, widely used in NMR structure elucidation, in that it exhibits simultaneously smoother and sharper energy surfaces, and results in protein structures with improved conformation, nonbonded atomic interactions, and accuracy.

Keywords: knowledge-based torsion angle potential, adaptive kernel density estimation, NMR protein structure calculation, protein structure validation.

Statement for the broader audience: Torsion angle probability densities extracted from a large database of amino acid crystallographic conformations were converted into a potential energy term to bias sampling during protein structure calculation towards highly populated regions in the database. Kernel density estimation was used to produce simultaneously smooth and sharp densities, which, when implemented as a potential in NMR structure computation, improved conformation, atomic packing, and accuracy, relative to a popular older potential.

Abbreviations: IIB^{Mtl}, cytoplasmic B domain of the mannitol transporter II^{mannitol}; BAF, barrier-to-autointegration factor; CNS, crystallography and NMR system; DHFR, apo dihydrofolate reductase; DinI, DNA damage inducible protein; EIN, N-terminal domain of enzyme I; GB1, B1 domain of protein G; KDE, kernel density estimation; KH3, C-terminal KH domain of heterogeneous nuclear ribonucleoprotein K; LM5-1, LM5-1 FYVE domain; NMR, nuclear magnetic resonance; NOE, nuclear Overhauser effect; PDB, protein data bank; RDC, residual dipolar coupling; RMS, root-mean-square; SrtA, sortease A in covalent complex with an LPXTG analog; Ubi, ubiquitin.

Introduction

Over the years, the accumulation of high-resolution X-ray structures in the Protein Data Bank (PDB)[1] has refined our knowledge of protein conformational preferences. Boundaries for the most favorable regions of the Ramachandran (ϕ, ψ) plot have shrunk,[2] and side chain rotamer distributions have become sharper and narrower than ever before.[3] Not only is this wealth of structural information exploited as validation criteria for newly generated models,[4] but also as a search bias during structure calculation. The latter approach aims at reproducing physically realistic conformational features of the structure database to alleviate the uncertainty associated with incomplete experimental information, such as that in low-resolution X-ray datasets. For instance, rotamer libraries can be used to fit side chain conformations to electron density,[5,6] and statistical potentials derived from database torsion angle distributions can supplement experimental restraints derived from NMR data. Driven by the relative sparseness of NMR data, the latter application was introduced more than 15 years ago,[7] and is at the center of the present study.

Statistical torsion angle potentials can be succinctly described as follows.[8] The probability density of torsion angles of interest is estimated from a database, and subsequently converted into potential energy by inversion of the Boltzmann equation. Thus (assuming a unit partition function as it cannot be directly obtained from the database),

$$E_a(\mathbf{x}) = -\beta \ln p(\mathbf{x}|a), \qquad (1)$$

where β is a constant, $\mathbf{x}$, one or more torsion angles, and $p(\mathbf{x}|a)$, the probability density of $\mathbf{x}$ given another variable a. $E_a(\mathbf{x})$ is a statistical potential that acts on $\mathbf{x}$ (given a). (It is also known as potential of mean force, and sometimes associated with other adjectives, such as empirical, database, and knowledge-based.) For example, if $\mathbf{x}$ consists of ϕ and ψ, and a is the amino acid residue type "alanine", then, during the calculation of a novel protein structure, $E_{\mathrm{Ala}}(\phi, \psi)$ biases the backbone torsion angles of alanines towards the densest regions of the Ramachandran distribution of alanine in the database. A collection of such potentials (e.g., one per residue type) is needed to handle every possible protein sequence.

Although statistical torsion angle potentials may be implemented under different conformational sampling techniques, NMR structure elucidation is typically achieved by molecular dynamics-based simulated annealing and gradient minimization, both of which call for smoothness and differentiability of the potentials. In this regard, the latest and most advanced implementation[9] relies on kernel density estimation (KDE) to obtain smooth, continuous probability densities involving all torsion angles within each residue type, the corresponding energy terms efficiently represented during structure calculations by cubic spline interpolation. However, since the main focus of that study was in structure prediction, NMR-relevant tests were reported as superficial, limited to the experimentally unrestrained minimization of previously solved NMR structures, omitting analysis of the accuracy of the resulting models, and their compatibility with the NMR data.

Possibly, the most thoroughly tested and most widely used statistical torsion angle potential in NMR is the so-called DELPHIC potential, developed by Kuszewski et al.,[7] which has evolved over time.[10-12] Included in the Xplor-NIH software package,[13,14] the latest version of this potential relies on a histogram-based approach for probability density estimation, the resulting energy (hyper)surfaces fit via an iterative protocol[11] whereby a quartic function is fit to the global minimum and subtracted from the surface. This procedure is repeated until a desired tolerance is met, and the energy surface is then represented by the sum of the quartic functions. Under comprehensive analysis,[15] the DELPHIC potential has been shown to significantly improve both structural validation criteria and accuracy, the former indicated by software tools such as WHAT IF,[16] the latter, by better agreement of the models with residual dipolar couplings (RDCs) purposely excluded from structure calculations (i.e., RDC cross-validation).

Despite the above-described encouraging results, visual inspection of the DELPHIC potential energy surfaces reveals roughness and other features (or lack thereof) that seem unsupported by torsion angle populations in modern databases (see below). Here, such deficiencies are addressed by a completely reformulated statistical torsion angle potential. A database of more than 10^6 quality-filtered residues is used to generate probability densities via adaptive kernel density estimation. This results in density estimates that are not only continuous and smooth overall, but also free of defects in regions of low density, where the noisy contribution of isolated points is automatically smoothed out. Finally, energy terms are efficiently represented during the course of structure calculation by cubic interpolation, from which forces are readily obtained. This new potential is incorporated into Xplor-NIH,[13,14] and tested on the structure calculation of ten proteins of various folds and sizes, using publicly available NMR restraints. The latter include RDCs, omitted from the calculations for cross-validation. The quality of backbone and side chain conformations, as well as that of nonbonded atomic interactions, was assessed using MolProbity[2,17,18] and WHAT IF.[16]

Results

The torsion angle database

The database on which the present study relies consists of 1,005,827 residues, extracted from protein crystal structures solved at a resolution of 1.8 Å or better, all atoms with B factors < 35 Å^2 and no serious atomic clashes reported by MolProbity. This database is a subset of the Top8000 database of almost 8,000 non-homologous protein chains (see Methods for details), kindly provided by Jane S. Richardson as a successor of the popular Top500 database.[2]

Residue type definitions and statistical approximations

The initial goal was to estimate the probability density function of all torsion angles within each residue type, starting from torsion angle instances in the database. Density estimates within a predefined grid were subsequently needed to obtain the corresponding energy values (via Equation 1), used in a cubic interpolation routine during structure computation (see Methods for details), where the statistical potential term (or terms; see below) was applied to all torsion angle degrees of freedom of residues with the corresponding type. Each residue was assigned only one type, following the order of precedence

Gly, cis-Pro, trans-Pro > prePro > Ala, Thr, Val, etc., (2) where cis/trans refers to the peptide bond conformation, prePro denotes a residue immediately preceding a proline in the primary structure, and the lowest-precedence types consist of all amino acid names minus Gly and Pro. For example, to estimate the (ϕ, ψ) probability density of Gly, $p(\phi, \psi|\text{Gly})$, all glycines in the database were used, regardless of whether they preceded a proline, whereas estimation of $p(\phi, \psi|\text{Ala})$ omitted pre-proline alanines. This residue classification is based on well-known distinctive torsion angle distributions (e.g., the relatively large, steric clash-free areas accessible to glycine, afforded by the lack of C^{β}), as well as the large size of the database. For example, whereas chemical similarity between tyrosine and phenylalanine has previously prompted their joint treatment to alleviate database scarcity (e.g., Ref. 10), here they yielded separate residue types, regardless of possible common features. Henceforth, Equation 2 will be implied whenever a name in it is used (e.g., “Ala” stands for “non-pre-proline alanine”).

Use of Equation 1 yields an energy term of the same dimensionality as the probability density function. Since all torsion angles within a residue type are involved, the highest possible dimensionality is six (e.g., Arg’s ϕ, ψ, χ_1, …, χ_4). However, the number of coefficients needed to represent the interpolated energy term during structure calculations becomes excessively computationally expensive beyond three dimensions.[9] Therefore, the problem is one of breaking probability densities of dimensionality > 3 into components of dimensionality ≤ 3, a statistical task that can be achieved by assuming conditional independence.[19] For example, in the case of Leu, χ_2 was assumed conditionally independent of ϕ and ψ given χ_1, which yields

$$p(\phi, \psi, \chi_1, \chi_2|\text{Leu}) = p(\phi, \psi, \chi_1|\text{Leu})\, p(\chi_1, \chi_2|\text{Leu}) \,/\, p(\chi_1|\text{Leu}). \qquad (3)$$

It should be noted that Equation 3 does not imply that χ_2 is independent of the backbone torsion angles, but that its dependence is indirect, via χ_1. Table I provides the probability density expressions for all residue types, including the statistical assumptions used to derive them. Such approximations rely on the number of atoms shared by torsion angles along the covalent framework of the residue: adjacent torsion angles (e.g., ϕ and χ_1) share more atoms than nonadjacent ones (e.g., ϕ and χ_2), and are, consequently, more likely to directly influence one another. Approximations similar to those in Table I have been previously used in another statistical torsion angle potential,[9] although the latter made more assumptions in that no three-dimensional probability densities were used for side chain torsion angles.

Two residue types in Table I deserve particular attention. First, the side chain torsion angles of Pro are highly correlated to one another due to covalent restrictions imposed by the ring. As a result, a single torsion angle, χ_2, can be used to determine the side chain conformation,[20] and the entire conformational space captured by $p(\phi, \psi, \chi_2|\text{Pro})$. Second, prePro is special in that it represents a diverse group (all non-glycine, non-proline residues immediately preceding a proline; Equation 2), with several residue subtypes. Due to insufficient pre-proline alanines ($\text{prePro}_{\text{Ala}}$ subtype) in the database, its (ϕ, ψ) density is represented by that of all prePro residues, $p(\phi, \psi|\text{prePro})$. The (ϕ, ψ, χ_1) distribution of the remaining prePro subtypes is captured by that of all prePro residues with at least χ_1, $p(\phi, \psi, \chi_1|\text{prePro})$. The distribution of torsion angles beyond χ_1 is assumed to be that of the corresponding residue, regardless of whether it precedes a proline. For example, for pre-proline leucines

$$p(\phi, \psi, \chi_1, \chi_2|\text{prePro}_{\text{Leu}}) = p(\phi, \psi, \chi_1|\text{prePro})\, p(\chi_2), \qquad (4)$$

where $p(\chi_2)$ is obtained from all leucines in the database. Although, Equation 4 implies the not-necessarily-true assertion that χ_2 is uncorrelated with the remaining torsion angles, such a correlation cannot be accurately obtained from the database due to the scarcity of pre-proline leucines, hence the current approximation. Similar considerations apply to the other prePro residue subtypes with torsion angles beyond χ_1 (see Table I). A more detailed version of Table I is provided as supplementary material (Supporting Information Table SI). It is noteworthy that while certain long-range correlations might be neglected by the above approximations, those based on atomic clashes are accounted for during structure calculations by repulsive interactions (see Methods for more details).

Adaptive kernel density estimation produces smooth, yet sharp potential surfaces

The general methodology chosen to extract the probability densities listed in Table I from the torsion angle database is kernel density estimation (KDE).[21] It consists of centering "bumps" or kernel functions on top of each database point; the density at any arbitrary position in torsion angle space is then estimated by summing the contribution of all kernels. In the present case, the kernels take the form of d-dimensional (d = 1, 2, or 3), symmetrical Gaussians, so that their overall smoothness is inherited by the density estimates. In particular, the adaptive version of KDE was used,[21] where the width of each Gaussian adapts to the local density in that narrow ones are placed in regions of high density, and wide ones in regions of low density. This has the effect of reproducing features at high local density, while smearing sparsely populated areas of torsion angle space (i.e., the "tails" of the distribution). The latter would appear bumpy if fixed-width kernels were used, and yield false high-energy local minima (via Equation 1) that could hamper structure calculations. This shortcoming of non-adaptive KDE may be mitigated by removal of isolated database points, followed by padding low-density regions with artificial points.[9] Here, the use of adaptive KDE rendered such modifications of the database unnecessary. Other examples of adaptive KDE in torsion angle space are provided elsewhere.[2,22]

The energy terms that comprise the new statistical torsion angle potential, henceforth referred to as “torsionDBPot”, follow directly from Table I and Equation 1. For example, Boltzmann inversion of Equation 3 yields

$$E_{Leu}(\phi, \psi, \chi_1, \chi_2) = E_{Leu}(\phi, \psi, \chi_1) + E_{Leu}(\chi_1, \chi_2) - E_{Leu}(\chi_1). \quad (5)$$

The subtraction of $E_{Leu}(\chi_1)$ intuitively accounts for the overweighting of χ_1 in the remainder of Equation 5 (note $\square_1$ appears in both $E_{Leu}(\phi, \psi, \chi_1)$ and $E_{Leu}(\chi_1, \chi_2)$). This is not an ad-hoc property of the potential, but one that arises naturally from the statistical treatment described in the previous section.

Figure 1 shows typical energy surfaces obtained from both the current version of the DELPHIC potential[12] in Xplor-NIH (Figures 1A and 1C) and torsionDBPot (Figures 1B and 1D), introduced in Xplor-NIH as part of the present work. Comparison of contour plots of the His (χ_1, χ_2) energy term reveals the absence of features in the DELPHIC potential (Figure 1A), notably, a shallow minimum at (62°, 83°) (corresponding to the sparsely populated p80° rotamer[3]), which is apparent in torsionDBPot (Figure 1B). Moreover, the DELPHIC potential suffers from noise and unrealistic shapes of energy surfaces, a problem exacerbated at high dimensions, as exemplified by the (ϕ, ψ, χ_1) energy term of Val (Figure 1C), which contrasts with the both smoother and sharper surfaces of torsionDBPot (Figures 1B and 1D). Visual inspection of several other surfaces indicates a prevalence of noise and multiple instances of missing features for the DELPHIC potential.

Effect of torsionDBPot on NMR structure calculations

The new statistical torsion angle potential, torsionDBPot, was tested on ten protein structures of diverse sizes (54 to 259 residues in length) and topologies (α, β, and αβ folds), listed in Table II. Structure calculations were performed with Xplor-NIH, enforcing publicly available NMR distance and torsion angle restraint sets. The latter reflect heterogeneity in their derivation from the experimental data. For example, whereas the backbone torsion angle restraints for the protein KH3[23] were obtained from scalar couplings, those for DinI[24] were derived from chemical shifts. The interpretation of nuclear Overhauser effects (NOEs) in terms of interproton distances also varies from one research group to another (six of which are represented here), a fact notably exemplified by SrtA,[25] which relied heavily on automation for NOE analysis, as opposed to, for example, IIB^{Mtl},[26] which followed a more conventional manual iterative approach. Differences in NOE data are also quantitative as the average number of long-range NOEs (i.e., between residues separated by more than five in the amino acid sequence) ranges from 1.8 to 10.6 per residue in the ten-protein set. Thus, the systems tested aim at representing a range of situations that may be encountered during the determination of a novel protein structure by NMR.

For each protein, three types of structure calculations were carried out, differing in the statistical torsion angle potential used: (i) none, (ii) the DELPHIC potential, or (iii) torsionDBPot. The generated structures were validated with MolProbity (Figure 2) and

WHAT IF (Figure 3), which report that calculations using either the DELPHIC or the torsionDBPot potential outperform in every criterion those that exclude such potentials. Furthermore, torsionDBPot improves the quality of the backbone conformation relative to the DELPHIC potential. Indeed, MolProbity indicates that, with the sole exception of EIN, the percent of Ramachandran outliers drops to its lowest values with torsionDBPot (Figure 2A), and the favored regions of the Ramachandran plot become more populated in every case, except for GB1, where a similar outcome is achieved with the DELPHIC potential (Figure 2B). These results agree with WHAT IF's Ramachandran plot appearance score, which improves throughout (Figure 3A) upon use of torsionDBPot. With regards to side chain conformation, both potentials perform similarly according to MolProbity: the percent of poor rotamers is slightly smaller with torsionDBPot for seven proteins, the remaining three yielding slightly better statistics with the DELPHIC potential (Figure 2C). On the other hand, WHAT IF favors torsionDBPot, which results in better χ_1/χ_2 rotamer normality scores for all proteins (Figure 3B).

Measures of quality of nonbonded atomic interactions, or atomic packing, are more independent validation criteria than those discussed above, in that they usually do not rely on the variables directly affected by the statistical torsion angle potentials, i.e., the torsion angles. Further, packing encompasses long-range features outside the scope of both the torsionDBPot and DELPHIC potentials, which only act at the local residue level. MolProbity's "clashscore",[27] the number of serious atomic overlaps per thousand atoms (see Methods for details), and WHAT IF's packing quality score,[28] which considers atomic distributions around different molecular fragments, are two such measures of packing. Relative to the DELPHIC potential, all structures generated with torsionDBPot display systematic improvements in the MolProbity clashscore, except for the slightly worse clashscore of IIB^{Mtl}, well within error bars (Figure 2D). This trend is also reflected by WHAT IF's packing quality score (Figure 3C).

The introduction of an additional energy term usually causes the agreement between calculated structures and terms in the original target function to deteriorate. It is therefore important to confirm that the improvements in conformation and atomic interactions afforded by torsionDBPot do not come at significant cost to the remaining terms, particularly those associated with the NMR data. Indeed, torsionDBPot is more compatible than the DELPHIC potential with the experimentally determined distances, as suggested by slightly lower root mean square (RMS) deviations from the upper and lower bounds of the distance restraints (Figure 4A). With regards to torsion angle restraints, RMS statistics suggest that some proteins exhibit better agreement when generated with the DELPHIC potential, others with torsionDBPot, but in every case the agreement is comparable to that of structures generated without either potential (except for ubiquitin whose publicly released restraints have unrealistically narrow bounds, hence the large RMS deviations) (Figure 4B).

The compatibility with experimental data excluded from structure calculations represents an independent test of structural accuracy. RDCs depend on the orientation of interatomic vectors relative to an external alignment tensor, and are commonly used for cross-validation.[29] Experimentally observed RDCs and those computed from the protein models were compared via an R-factor,[30] which ranges from 0% (perfect correlation) to 100% (no correlation). Figure 4C shows that inclusion of the DELPHIC potential in structure calculations significantly improves the fit to RDCs, as previously reported elsewhere.[15] torsionDBPot improves the fit even further in all cases, with the exception of IIB$^{\mathrm{Mtl}}$, where a slightly worse fit is observed.

Discussion

The previously developed statistical torsion angle potential[7,10-12] in Xplor-NIH[13,14] (or its precursors, X-PLOR[31] and CNS[32]) has become an important tool in NMR protein structure determination, inspiring similar implementations in other software packages.[9,33,34] Although originally applied to solution NMR, the statistical torsion angle potential in Xplor-NIH (named DELPHIC) has additionally made significant contributions with other types of experimental data, such as combination of solution and solid state NMR,[35-37] combination of solution NMR and small- and wide-angle X-ray scattering,[38,39] combination of solid state NMR and X-ray diffraction,[40] and purely solid state NMR data (e.g., Refs. 41-43). With this in mind, a "First, do no harm" approach was followed in the development of torsionDBPot, a new statistical torsion angle potential in Xplor-NIH. (Note that the DELPHIC potential remains available for backwards compatibility.) Indeed, while accomplishing similar or better fit to experimental restraints relative to the DELPHIC potential, torsionDBPot improves the quality of protein conformation and nonbonded atomic interactions. This is summarized by the overall MolProbity score[18] (the lower the better), which improves in every case tested (Figure 2E). Moreover, such benefits are concomitant with enhanced structural accuracy, as suggested by better agreement with cross-validated RDCs. Albeit relatively moderate, the above-reported improvements are consistent across the entire protein test set.

Despite their usefulness in protein structure prediction and experiment-based determination, a general concern with statistical potentials of any kind is that their inherent average nature may bias structures away from features that are real although poorly represented in the database. Use of experimental data, however, as in the present case, ameliorates this bias because the data are allowed to trump the statistical potential whenever possible. Thus, a rarely observed conformation firmly supported by NMR restraints should prevail over torsionDBPot in structure calculations (otherwise, restraint violations would arise highlighting the unusual region). In this light, flexibly disordered regions appear problematic as they are both poorly represented in the structured, low-B factor database, and usually sparsely restrained by NMR data Future work may address the need for alternate descriptions of nonregular regions. For the ten proteins studied here the current protocol is clearly a step forward in generating high-quality NMR structures.

Methods

The torsion angle database

The starting point for the compilation of the torsion angle database used in this study is the Top8000 database (kindly provided by Jane S. Richardson, Duke University), of almost 8,000 chains with X-ray structure resolution better than 2.0 Å, less than 70% sequence identity, and other satisfied filters, notably: chain MolProbity score < 2.0, ≤ 5% of residues with bond lengths and angles outside 4 standard deviations from standard geometry, ≤ 5% of residues with C^β deviations > 0.25 Å, and best average of resolution and MolProbity score among the 70% homology cluster the chain represents. In addition, similar to previous versions of this database,[2,3] the Top8000 contains flipped planar side chain terminal groups of asparagines, glutamines, and histidines, when justified by analysis of atomic clashes and H-bonding.[3] Further details are provided at the Richardson Lab's website (http://kinemage.biochem.duke.edu/databases/top8000.php). The Top8000 database was obtained as a table, each row containing information on a single residue, such as its torsion angles, resolution, atomic clashes (if any), etc.

As discussed in the Results section, the Top8000 database was subjected to more stringent filters to generate the custom database used by our new statistical torsion angle potential, torsionDBPot. Specifically, only chains with X-ray resolution of 1.8 Å or better were considered, residues from which were included in the custom database only if all their atoms had B factors < 35 $Å^2$ and no serious clashes reported by MolProbity.[2,17,18] Moreover, leucines with (χ_1, χ_2) pairs within regions that represent misfit rotamers[3,10] were avoided; a total of 553 such misfits were encountered and removed after resolution, B factor, and clash filtering. The resulting torsion angle database contains 1,005,827 residues.

A new Python module, torsionDBTools, has been added to Xplor-NIH to facilitate the extraction of torsion angles from Cartesian coordinates (i.e., PDB files). Although thoroughly tested, this module was not used here, as the Top8000 database already provided torsion angle, along with other useful information (see above). However, the module should prove useful in the derivation of new statistical torsion angle potentials from arbitrary subsets of the PDB (e.g., coil databases).

Torsion angle probability densities via adaptive kernel density estimation

The goal is to accurately estimate the probability density function of torsion angles of interest from the database, with the additional requirement that the estimate be smooth. The torsion angles under consideration can be represented by a column vector **x**, which defines a *d*-dimensional space where *n* database points $\mathbf{X}_{1}$, …, $\mathbf{X}_{n}$ are found (e.g., if $\mathbf{x}^{\mathrm{T}} = (\phi, \psi)$, $d = 2$, where T denotes vector transpose). A first approximation towards extracting the probability density from the database is to perform kernel density estimation (KDE), by summing over "bumps" or kernels centered at the observed database points.[21] KDE with a symmetrical Gaussian kernel function and window width *h* is defined by

$$\tilde{p}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{N}(\boldsymbol{\mu} = \mathbf{X}_i, \boldsymbol{\Sigma} = h^2\mathbf{I}), \tag{6}$$

where the N-notation is used for the *d*-dimensional (or *d*-variate) Gaussian, with mean vector μ and covariance matrix Σ. (**I** is the identity matrix). Explicitly,

$$\mathrm{N}(\boldsymbol{\mu} = \mathbf{X}_i, \boldsymbol{\Sigma} = h^2\mathbf{I}) = \frac{1}{(2\pi h^2)^{d/2}} \exp\left(-\frac{1}{2h^2}(\mathbf{x} - \mathbf{X}_i)^{\mathrm{T}}(\mathbf{x} - \mathbf{X}_i)\right). \tag{7}$$

In one dimension, for example, the left-hand side of Equation 7 simply becomes $\mathrm{N}(\mu = \mathrm{X}_i, \sigma^2 = h^2)$, where the boldface vector/matrix notation is no longer necessary, and the variance, σ^2, replaces the covariance matrix. It is noteworthy that, for the sake of simplicity, all probability density functions in this section (including Equations 6 and 7 above) tacitly imply the residue type conditional, explicit elsewhere in the text (cf. Equation 1).

Gaussians are nonnegative and integrate to one, leading $\tilde{p}(\mathbf{x})$ to also be nonnegative and integrate to one, as any probability density function must. Also, since Gaussians are continuous and smooth, so is $\tilde{p}(\mathbf{x})$. The degree of smoothing is additionally controlled by the choice of window width *h* (i.e., σ, the standard deviation in the one-dimensional case). Throughout our work, the periodicity of angular values is dealt with by augmenting the database with shifted copies of the original.[21] Following the one-dimensional example, if the torsion angle under study is defined in the interval [−180°, 180°), adding copies of the database at intervals [−540°, −180°) and [180°, 540°) results in a new database $\{X_1 - 360°, \ldots, X_n - 360°, X_1, \ldots, X_n, X_1 + 360°, \ldots, X_n + 360°\}$. Performing KDE on this augmented database with Equation 6 (where *n* is still the original number of points) accounts for the boundary condition.

Despite its obvious advantages over simpler density estimation methods such as the histogram, KDE has the tendency to produce noise in regions of low density, arising from individual, isolated bumps, a problem exacerbated in high dimensions. Here, the solution chosen was the use of kernels with variable window width—as opposed to the fixed-width kernels of Equation 6—so that narrow kernels are placed in regions of high density, and wide ones in regions of low density. This method is called adaptive KDE,[21] as the window width adapts to the local density, which is preliminary estimated via standard KDE (Equation 6, in this context usually referred to as the pilot estimate). Using again symmetrical Gaussians, adaptive KDE takes the mathematical form

$$p(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{N}(\boldsymbol{\mu} = \mathbf{X}_i, \boldsymbol{\Sigma}_i = (h\lambda_i)^2 \mathbf{I}), \qquad (8)$$

where local bandwidth factors λ_i are given by

$$\lambda_i = \left(\frac{g}{\tilde{p}(\mathbf{X}_i)} \right)^{\alpha} . \tag{9}$$

In Equation 9, $g = \left(\prod_{i=1}^{n} \tilde{p}(\mathbf{X}_i) \right)^{1/n}$, the geometric mean of the $\tilde{p}(\mathbf{X}_i)$ (a constant), and α is set to 0.5 as recommended elsewhere.[21] The formulas for the variable-width, symmetrical Gaussians in Equation 8 can be readily obtained from Equation 7 by replacing h by $h\lambda_i$. Once the λ_i's are determined for a joint probability density estimate (e.g., $p(\phi, \psi. \chi_1)$), the marginal probability density estimate of one (or more) torsion angles (e.g., $p(\chi_1)$) can be computed in a straightforward manner. When more than one joint probability density estimate is available, the marginal probability density of a common torsion angle is extracted from each joint distribution, and an average density computed.

In the present study, adaptive KDE was performed as described above, for one, two, and three dimensions, where h was given the values of 4°, 5°, and 6°, respectively—as prescribed elsewhere,[21] h takes the same value in Equations 6 and 8. All calculations were performed using the Python module densityEstimation implemented within Xplor-NIH[13,14] for the present purposes.

Cubic interpolation of energy terms

The different energy terms that stem from the application of Equation 1 to the adaptive KDE-based probability density functions in Table I (see Supporting Information Table SI for a more detailed version of Table I) were evaluated on a grid used for cubic interpolation with periodic boundary conditions. In one and two dimensions, a uniform grid with 10°-spacing was used. Extending this strategy to the construction of a three-dimensional grid results in an unacceptable increase of computer memory requirements.

Consequently, a non-uniform grid was devised with 10°-spacing around each energy minimum, and wider spacing elsewhere. Specifically, the axis along one of the three dimensions is uniformly marked every 10°, and a tick mark retained only if within a distance *r* of the coordinate of a minimum along that dimension. Subsequently, to avoid under-sampling, if two adjacent tick marks are farther apart than 10°, a new one is added equidistant from the two. The same procedure is performed with the axes along the remaining two dimensions, and the grid constructed with the three sparsely sampled axes. $r = 23°$ for all residue types (Equation 2), with the exception of cis/trans-Pro, where $r = 30°$, the denser sampling afforded by the fact that the minima are confined to a small region of torsion angle space.

Within Xplor-NIH, cubic interpolation routines in one and two dimensions[44] were already present, and were previously exploited by another potential term.[41] Three-dimensional cubic interpolation capabilities, as described elsewhere,[45] were added to Xplor-NIH (spline3D Python module) for the present purposes, and have already been successfully applied to a recent unrelated problem.[46] The interpolated energy terms make up the new statistical torsion angle potential, torsionDBPot, which is set up with the newly added module torsionDBPotTools. In addition, torsionDBPotTools contains the auxiliary function find_minima, which characterizes a queried torsionDBPot surface by listing the number, location, and depth of all its minima, and is useful for the comparison of different surfaces.

Structure calculations

Structures were calculated with Xplor-NIH,[13,14] using two conventional simulated annealing protocols: a first one for folding an initially extended conformation, and a second one for the subsequent refinement of a selected folded model. Both protocols, based on the internal variable module,[47] share the same basic scheme, comprising the following stages (respecting their order during calculations): (i) high-temperature torsion angle dynamics (3,500 K for folding, 3,000 K for refinement), the smallest of 15 ps or 15,000 timesteps in length, subject to torsion angle restraints (k_{ta} = 10 kcal mol^{-1} rad^{-2}), distance restraints (k_{dist} = 2 kcal mol^{-1} Å^{-2}), and van der Waals-like repulsions[48] (k_{vdw} = 0.004 kcal mol^{-1} Å^{-4}; only C$^{\alpha}$–C$^{\alpha}$ interactions active, with a van der Waals radius scale factor s_{vdw} = 1.2), where $k_{\square}$ represents the force constant of energy term η; (ii) torsion angle dynamics with simulated annealing, where the temperature is reduced from the initial value (see above) to 25 K in steps of 12.5 K (the smallest of 0.2 ps/step or 200 timesteps/step for folding, the smallest of 0.63 ps/step or 630 timesteps/step for refinement), k_{ta} = 200 kcal mol^{-1} rad^{-2}, and k_{dist}, k_{vdw}, and s_{vdw} are geometrically increased from 2 to 30 kcal mol^{-1} Å^{-2}, 0.004 to 4 kcal mol^{-1} Å^{-4}, and 0.9 to 0.8, respectively (all van der Waals interactions active (see exceptions below) in this stage, a feature maintained in subsequent stages, as well as the final values of force constants and s_{vdw}); (iii) 500 steps of Powell torsion angle minimization; (iv) 500 steps of Powell Cartesian minimization. When including either the DELPHIC[12] or the torsionDBPot statistical torsion angle potential term, its force constant is set to 0.002 kcal mol^{-1} rad^{-2} in stage (i), from which it geometrically increases to 1 (DELPHIC) or 2 kcal mol^{-1} rad^{-2} (torsionDBPot) in stage (ii), values maintained until the end of the protocol. Although many steric interactions are already accounted for by both the DELPHIC and torsionDBPot potentials (e.g., those in eclipsed conformations that result in staggered side chain rotamer distributions), the prevention of any atomic overlap is an essential part of any force field. Here, a compromise is achieved by allowing repulsions only between atoms separated by more than three covalent bonds whenever the DELPHIC or torsionDBPot potentials are used; when they are not used, repulsions are allowed only between atoms separated by more than two covalent bonds.

The folding protocol generated 100 structures, from which the one with the lowest experimental energy (i.e., energy from distance and torsion angle restraints) was selected for refinement. The refinement protocol generated 100 structures, and the top 20 ranked by the experimental energy were selected for further analysis. Computer memory requirements for torsionDBPot were similar to those of the DELPHIC potential.

Experimentally determined distance, torsion angle, and RDC restraints (the latter excluded from the structure calculations, and used only for purposes of cross-validation; see below for details) were obtained from the PDB for the ten proteins listed in Table II.

Structure validation

The quality of backbone and side chain conformations, as well as that of nonbonded interatomic interactions in the calculated protein structures were assessed with MolProbity[2,17,18] and WHAT IF.[16] The increase of a WHAT IF score was considered an improvement of the associated quality criterion (a proper “score” behavior). On the other hand, MolProbity’s overall score[18] and clashscore[27] are actually costs whose decrease reflect improvement. It is noteworthy that the clashscore (number of serious atomic overlaps per thousand atoms) ignores clashes between pairs of heavy atoms within three or fewer covalent bonds, and between pairs of atoms where one or both are hydrogens within four of fewer bonds. In other words, the MolProbity clashscore ignores clashes between atoms whose relative positions are directly affected by the statistical torsion angle potentials during structure calculations, thus making it a more independent measure of structure quality as opposed to, for example, the percentage of poor side chain rotamers.

Agreement between structures and residual dipolar couplings

RDCs were fit to calculated structures by singular value decomposition[49] with Xplor-NIH, which additionally reports the R-factor measure of fit,[30]

$$R = 100\sqrt{\frac{5\left\langle \left(D_{obs}^{AB} - D_{calc}^{AB}\right)^2 \right\rangle}{2\left(D_a^{AB}\right)^2\left(4 + 3Rh^2\right)}}, \tag{10}$$

where D_{obs}^{AB} is the experimentally observed and D_{calc}^{AB} the structure-calculated RDC for nuclei pair type A–B (e.g., $^1H^N$–^{15}N) in a given molecular alignment medium, D_a^{AB} and *Rh* are the axial component and the rhombicity of the alignment tensor, respectively, and angular brackets denote averaging over the entire A–B RDC dataset. A single unweighted R-factor average over all nuclei pair types and media was used to assess the overall fit.

Availability

The new statistical torsion angle potential, torsionDBPot, is part of the Xplor-NIH software suite (version 2.31), downloadable from the web (http://nmr.cit.nih.gov/xplor-nih/).

Acknowledgements

Jane S. Richardson provided the Top8000 database, and, along with Daniel A. Keedy and Vincent B. Chen, participated in useful discussions. Robert T. Clubb provided Xplor-NIH topology and parameter files for Sortease A; Nuttee Suree and Albert Chan helped with their setup. This work was supported by the NIH Intramural Research Programs of CIT (C.D.S. and G.A.B.) and NIDDK (G.M.C.), and by the AIDS Targeted Antiviral Program of the Office of the Director of the NIH (G. M. C.).

Competing financial interests: The authors declare no competing financial interest.

References

1. Berman H, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank. Nat Struct Biol 10:980-980.
2. Lovell SC, Davis IW, Adrendall WB, de Bakker PIW, Word JM, Prisant MG, Richardson JS, Richardson DC (2003) Structure validation by Cα geometry: φ, ψ and Cβ deviation. Proteins 50:437-450.
3. Lovell SC, Word JM, Richardson JS, Richardson DC (2000) The penultimate rotamer library. Proteins 40:389-408.
4. Read RJ, Adams PD, Arendall WB, III, Brunger AT, Emsley P, Joosten RP, Kleywegt GJ, Krissinel EB, Lutteke T, Otwinowski Z and others (2011) A new

generation of crystallographic validation tools for the protein data bank. Structure 19:1395-412.

5. Adams PD, Grosse-Kunstleve RW, Hung LW, Ioerger TR, McCoy AJ, Moriarty NW, Read RJ, Sacchettini JC, Sauter NK, Terwilliger TC (2002) PHENIX: building new software for automated crystallographic structure determination. Acta Crystallogr D Biol Crystallogr 58:1948-1954.

6. Headd JJ, Immormino RM, Keedy DA, Emsley P, Richardson DC, Richardson JS (2009) Autofix for backward-fit sidechains: using MolProbity and real-space refinement to put misfits in their place. J Struct Funct Genomics 10:83-93.

7. Kuszewski J, Gronenborn AM, Clore GM (1996) Improving the quality of NMR and crystallographic protein structures by means of a conformational database potential derived from structure databases. Protein Sci 5:1067-1080.

8. Sippl MJ (1990) Calculation of conformational ensembles from potentials of mean force - an approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 213:859-883.

9. Amir ED, Kalisman N, Keasar C (2008) Differentiable, multi-dimensional, knowledge-based energy terms for torsion angle probabilities and propensities. Proteins 72:62-73.

10. Kuszewski J, Gronenborn AM, Clore GM (1997) Improvements and extensions in the conformational database potential for the refinement of NMR and X-ray structures of proteins and nucleic acids. J Magn Reson 125:171-177.

11. Kuszewski J, Clore GM (2000) Sources of and solutions to problems in the refinement of protein NMR structures against torsion angle potentials of mean force. J Magn Reson 146:249-254.

12. Clore GM, Kuszewski J (2002) χ1 Rotamer populations and angles of mobile surface side chains are accurately predicted by a torsion angle database potential of mean force. J Am Chem Soc 124:2866-2867.

13. Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. J Magn Reson 160:65-73.

14. Schwieters CD, Kuszewski JJ, Clore GM (2006) Using Xplor-NIH for NMR molecular structure determination. Prog Nucl Mag Res Sp 48:47-62.

15. Mertens HDT, Gooley PR (2005) Validating the use of database potentials in protein structure determination by NMR. FEBS Lett 579:5542-5548.

16. Vriend G (1990) What If - a molecular modeling and drug design program. J Mol Graph 8:52-56.

17. Davis IW, Leaver-Fay A, Chen VB, Block JN, Kapral GJ, Wang X, Murray LW, Arendall WB, Snoeyink J, Richardson JS and others (2007) MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. Nucleic Acids Res 35:W375-W383.

18. Chen VB, Arendall WB, Headd JJ, Keedy DA, Immormino RM, Kapral GJ, Murray LW, Richardson JS, Richardson DC (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Cryst D66:12-21.

19. Russell SJ, Norvig P (2003) Artificial Intelligence: a Modern Approach, Prentice Hall/Pearson Education, Upper Saddle River, N.J.

20. DeTar DF, Luthra NP (1977) Conformations of proline. J Am Chem Soc 99:1232-1244.

21. Silverman BW (1986) Density estimation for statistics and data analysis, Chapman and Hall, London, New York.

22. Shapovalov MV, Dunbrack RL (2011) A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure 19:844-858.

23. Baber JL, Libutti D, Levens D, Tjandra N (1999) High precision solution structure of the C-terminal KH domain of heterogeneous nuclear ribonucleoprotein K, a c-myc transcription factor. J Mol Biol 289:949-962.

24. Ramirez BE, Voloshin ON, Camerini-Otero RD, Bax A (2000) Solution structure of DinI provides insight into its mode of RecA inactivation. Protein Sci 9:2161-2169.

25. Suree N, Liew CK, Villareal VA, Thieu W, Fadeev EA, Clemens JJ, Jung ME, Clubb RT (2009) The structure of the Staphylococcus aureus sortase-substrate complex reveals how the universally conserved LPXTG sorting signal is recognized. J Biol Chem 284:24465-24477.

26. Legler PM, Cai ML, Peterkofsky A, Clore GM (2004) Three-dimensional solution structure of the cytoplasmic B domain of the mannitol transporter II$^{\mathrm{Mannitol}}$ of the Escherichia coli phosphotransferase system. J Biol Chem 279:39115-39121.

27. Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME, Presley BK, Richardson JS, Richardson DC (1999) Visualizing and quantifying molecular goodness-of-fit:

Small-probe contact dots with explicit hydrogen atoms. J Mol Biol 285:1711-1733.

28. Vriend G, Sander C (1993) Quality control of protein models: Directional atomic contact analysis. J Appl Cryst 26:47-60.

29. Bax A, Grishaev A (2005) Weak alignment NMR: a hawk-eyed view of biomolecular structure. Curr Opin Struct Biol 15:563-570.

30. Clore GM, Garrett DS (1999) R-factor, free R, and complete cross-validation for dipolar coupling refinement of NMR structures. J Am Chem Soc 121:9008-9012.

31. Brünger AT (1992) X-PLOR, Version 3.1: a system for X-ray crystallography and NMR, Yale University Press, New Haven.

32. Brünger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS and others (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 54:905-921.

33. Bertini I, Cavallaro G, Luchinat C, Poli I (2003) A use of Ramachandran potentials in protein solution structure determinations. J Biomol NMR 26:355-66.

34. Yang JS, Kim JH, Oh S, Han G, Lee S, Lee J (2012) STAP refinement of the NMR database: a database of 2405 refined solution NMR structures. Nucleic Acids Res 40:D525-D530.

35. Shi L, Traaseth NJ, Verardi R, Cembran A, Gao JL, Veglia G (2009) A refinement protocol to determine structure, topology, and depth of insertion of membrane proteins using hybrid solution and solid-state NMR restraints. J Biomol NMR 44:195-205.

36. Traaseth NJ, Shi L, Verardi R, Mullen DG, Barany G, Veglia G (2009) Structure and topology of monomeric phospholamban in lipid membranes determined by a hybrid solution and solid-state NMR approach. Proc Natl Acad Sci USA 106:10165-10170.

37. Verardi R, Shi L, Traaseth NJ, Walsh N, Veglia G (2011) Structural topology of phospholamban pentamer in lipid bilayers by a hybrid solution and solid-state NMR method. Proc Natl Acad Sci USA 108:9101-9106.

38. Schwieters CD, Suh JY, Grishaev A, Ghirlando R, Takayama Y, Clore GM (2010) Solution structure of the 128 kDa enzyme I dimer from Escherichia coli and its 146 kDa complex with HPr using residual dipolar couplings and small- and wide-angle X-ray scattering. J Am Chem Soc 132:13026-13045.

39. Takayama Y, Schwieters CD, Grishaev A, Ghirlando R, Clore GM (2011) Combined use of residual dipolar couplings and solution X-ray scattering to rapidly probe rigid-body conformational transitions in a non-phosphorylatable active-site mutant of the 128 kDa enzyme I dimer. J Am Chem Soc 133:424-427.

40. Tang M, Sperling LJ, Berthold DA, Schwieters CD, Nesbitt AE, Nieuwkoop AJ, Gennis RB, Rienstra CM (2011) High-resolution membrane protein structure by joint calculations with solid-state NMR and X-ray experimental data. J Biomol NMR 51:227-233.

41. Wylie BJ, Schwieters CD, Oldfield E, Rienstra CM (2009) Protein structure refinement using 13Cα chemical shift tensors. J Am Chem Soc 131:985-992.

42. Sengupta I, Nadaud PS, Helmus JJ, Schwieters CD, Jaroniec CP (2012) Protein fold determined by paramagnetic magic-angle spinning solid-state NMR spectroscopy. Nat Chem 4:410-7.

43. Tian Y, Schwieters CD, Opella SJ, Marassi FM (2012) AssignFit: A program for simultaneous assignment and structure refinement from solid-state NMR spectra. J Magn Reson 214:42-50.

44. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (2007) Numerical recipes: the art of scientific computing, Cambridge University Press, Cambridge, New York.

45. Lekien F, Marsden J (2005) Tricubic interpolation in three dimensions. Int J Numer Meth Eng 63:455-471.

46. Hu KN, Qiang W, Bermejo GA, Schwieters CD, Tycko R (2012) Restraints on backbone conformations in solid state NMR studies of uniformly labeled proteins from quantitative amide ^{15}N-^{15}N and carbonyl ^{13}C-^{13}C dipolar recoupling data. J Magn Reson 218:115-27.

47. Schwieters CD, Clore GM (2001) Internal coordinates for molecular dynamics and minimization in structure determination and refinement. J Magn Reson 152:288-302.

48. Nilges M, Clore GM, Gronenborn AM (1988) Determination of 3-dimensional structures of proteins from interproton distance data by hybrid distance geometry-dynamical simulated annealing calculations. FEBS Lett 229:317-324.

49. Losonczi JA, Andrec M, Fischer MW, Prestegard JH (1999) Order matrix analysis of residual dipolar couplings using singular value decomposition. J Magn Reson 138:334-42.

50. Hunter JD (2007) Matplotlib: A 2D graphics environment. Comput Sci Eng 9:90-95.

51. Ramachandran P, Varoquaux G (2011) Mayavi: 3D visualization of scientific data. Comput Sci Eng 13:40-50.

52. Kuszewski J, Gronenborn AM, Clore GM (1999) Improving the packing and accuracy of NMR structures with a pseudopotential for the radius of gyration. J Am Chem Soc 121:2337-2338.

53. Cornilescu G, Marquardt JL, Ottiger M, Bax A (1998) Validation of protein structure from anisotropic carbonyl chemical shifts in a dilute liquid crystalline phase. J Am Chem Soc 120:6836-6837.

54. Mertens HDT, Callaghan JM, Swarbrick JD, Mcconville MJ, Gooley PR (2007) A high-resolution solution structure of a trypanosomatid FYVE domain. Protein Sci 16:2552-2559.

55. Cai M, Huang Y, Zheng R, Wei SQ, Ghirlando R, Lee MS, Craigie R, Gronenborn AM, Clore GM (1998) Solution structure of the cellular factor BAF responsible for protecting retroviral DNA from autointegration. Nat Struct Biol 5:903-909.

56. Feeney J, Birdsall B, Kovalevskaya NV, Smurnyy YD, Peran EMN, Polshakov VI (2011) NMR structures of Apo L. casei dihydrofolate reductase and its complexes with trimethoprim and NADPH: Contributions to positive cooperative

binding from ligand-induced refolding, conformational changes, and interligand hydrophobic interactions. Biochemistry 50:3609-3620.

57. Garrett DS, Seok YJ, Liao DI, Peterkofsky A, Gronenborn AM, Clore GM (1997) Solution structure of the 30 kDa N-terminal domain of enzyme I of the Escherichia coli phosphoenolpyruvate:sugar phosphotransferase system by multidimensional NMR. Biochemistry 36:2517-2530.

58. Garrett DS, Seok YJ, Peterkofsky A, Gronenborn AM, Clore GM (1999) Solution structure of the 40,000 M-r phosphoryl transfer complex between the N-terminal domain of enzyme I and HPr. Nat Struct Biol 6:166-173.

Table I. Probability Density Expressions Extracted from the Torsion Angle Database

Residue Type	Probability Density Function [a]	Statistical Approximation [b]
Gly, Ala	$p(\phi,\psi)$	None
Thr, Val, Ser, Cys	$p(\phi,\psi,\chi_1)$	None
cis-Pro, trans-Pro	$p(\phi,\psi,\chi_2)$	χ_2 determines other ring torsion angles
Asp, Asn, Ile, Leu, His, Trp, Tyr, Phe	$\dfrac{p(\phi,\psi,\chi_1)p(\chi_1,\chi_2)}{p(\chi_1)}$	$\chi_2 \perp \phi,\psi$ given χ_1
Met, Glu, Gln	$\dfrac{p(\phi,\psi,\chi_1)p(\chi_1,\chi_2,\chi_3)}{p(\chi_1)}$	$\begin{cases}\chi_2 \perp \phi,\psi \text{ given } \chi_1 \\ \chi_3 \perp \phi,\psi \text{ given } \chi_1,\chi_2\end{cases}$
Lys, Arg	$\dfrac{p(\phi,\psi,\chi_1)p(\chi_1,\chi_2,\chi_3)p(\chi_2,\chi_3,\chi_4)}{p(\chi_1)p(\chi_2,\chi_3)}$	$\begin{cases}\chi_2 \perp \phi,\psi \text{ given } \chi_1 \\ \chi_3 \perp \phi,\psi \text{ given } \chi_1,\chi_2 \\ \chi_4 \perp \phi,\psi \text{ given } \chi_2,\chi_3\end{cases}$
prePro	$p(\phi,\psi \mid \text{prePro})$ $p(\phi,\psi,\chi_1 \mid \text{prePro})$ $p(\phi,\psi,\chi_1 \mid \text{prePro})p(\chi_2)$ $p(\phi,\psi,\chi_1 \mid \text{prePro})p(\chi_2,\chi_3)$ $p(\phi,\psi,\chi_1 \mid \text{prePro})p(\chi_2,\chi_3,\chi_4)$	None None $\chi_2 \perp \phi,\psi,\chi_1$ $\chi_2,\chi_3 \perp \phi,\psi,\chi_1$ $\chi_2,\chi_3,\chi_4 \perp \phi,\psi,\chi_1$

[a] For simplicity, probability density functions omit conditionals whenever possible. For example, for residue types Gly and Ala, $p(\phi, \psi)$ implies $p(\phi, \psi|\text{Gly})$ and $p(\phi, \psi|\text{Ala})$, respectively. In the case of prePro, the explicit use of a conditional is informative. For example, the probability density function of pre-proline arginines is represented by $p(\phi, \psi. \chi_1|\text{prePro})$ $p(\chi_2, \chi_3, \chi_4)$, where $p(\phi, \psi. \chi_1|\text{prePro})$ corresponds to all non-glycine,

non-proline pre-proline residues (with at least one side chain torsion angle), and $p(\chi_2, \chi_3, \chi_4)$ to all arginines, regardless whether they precede a proline (see text for another example).

[b] The statistical approximations used to arrive at the different probability density expressions are indicated, where the orthogonality sign ($\perp$) means “independent of”.

Table II. Proteins Used in Test Structure Calculations

Protein (Name and Abbreviation)	Residues	Fold	NMR Data (PDB ID)	References
B1 domain of protein G (GB1)	54	αβ	3GB1	52
Ubiquitin (Ubi)	76	αβ	1D3Z	53
DNA damage inducible protein 1 (DinI)	81	αβ	1GHH	24
LM5-1 FYVE domain (LM5-1)	84	αβ	1Z2Q	54
C-terminal KH domain of heterogeneous nuclear ribonucleoprotein K (KH3)	89	αβ	1KHM	23
Barrier-to-autointegration factor (BAF, chain A only)	89	α	2EZX	55
Cytoplasmic B domain of the mannitol transporter II^{mannitol} (IIB^{Mtl})	97	αβ	1VKR	26
Sortease A in covalent complex with an LPXTG analog (SrtA)	148	β	2KID	25
Apo dihydrofolate reductase (DHFR)	162	αβ	2L28	56
N-terminal domain of enzyme I (EIN)	259	αβ	1EZA (distance, torsion angle) 3EZA (RDC)	57,58

Figure Legends

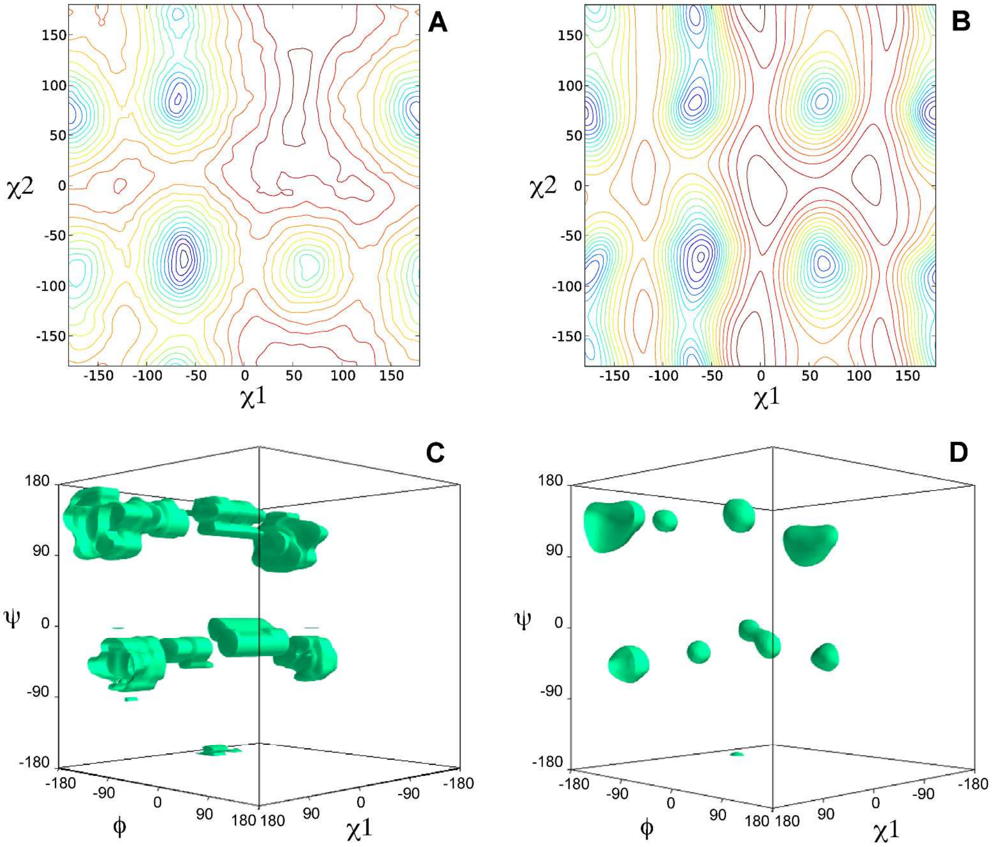
Figure 1. Representative energy surfaces of the DELPHIC and torsionDBPot statistical torsion angle potentials in Xplor-NIH. (A and B) Contour plots of the His (χ_1, χ_2) energy term in the DELPHIC potential and torsionDBPot, respectively. (C and D) Single isoenergetic surfaces of the Val (ϕ, ψ, χ_1) energy term in the DELPHIC potential and torsionDBPot, respectively. Panels A and B were generated with Matplotlib,[50] C and D with Mayavi.[51] All units are in degrees.

Figure 2. MolProbity validation. Each barplot displays a Molprobity validation statistic for structure ensembles of different proteins, with bars representing the mean ± standard deviation computed from 20 structures. Structure calculations without any statistical torsion angle potential (black), with the DELPHIC potential[12] (gray), and with the new torsionDBPot potential (white) are included. Abbreviated protein names are used; for full names see Table II. The clashscore[27] (panel D) and the MolProbity score[18] (panel E) are costs: the lower the better. Barplots in this and all other figures were generated with Matplotlib.[50]

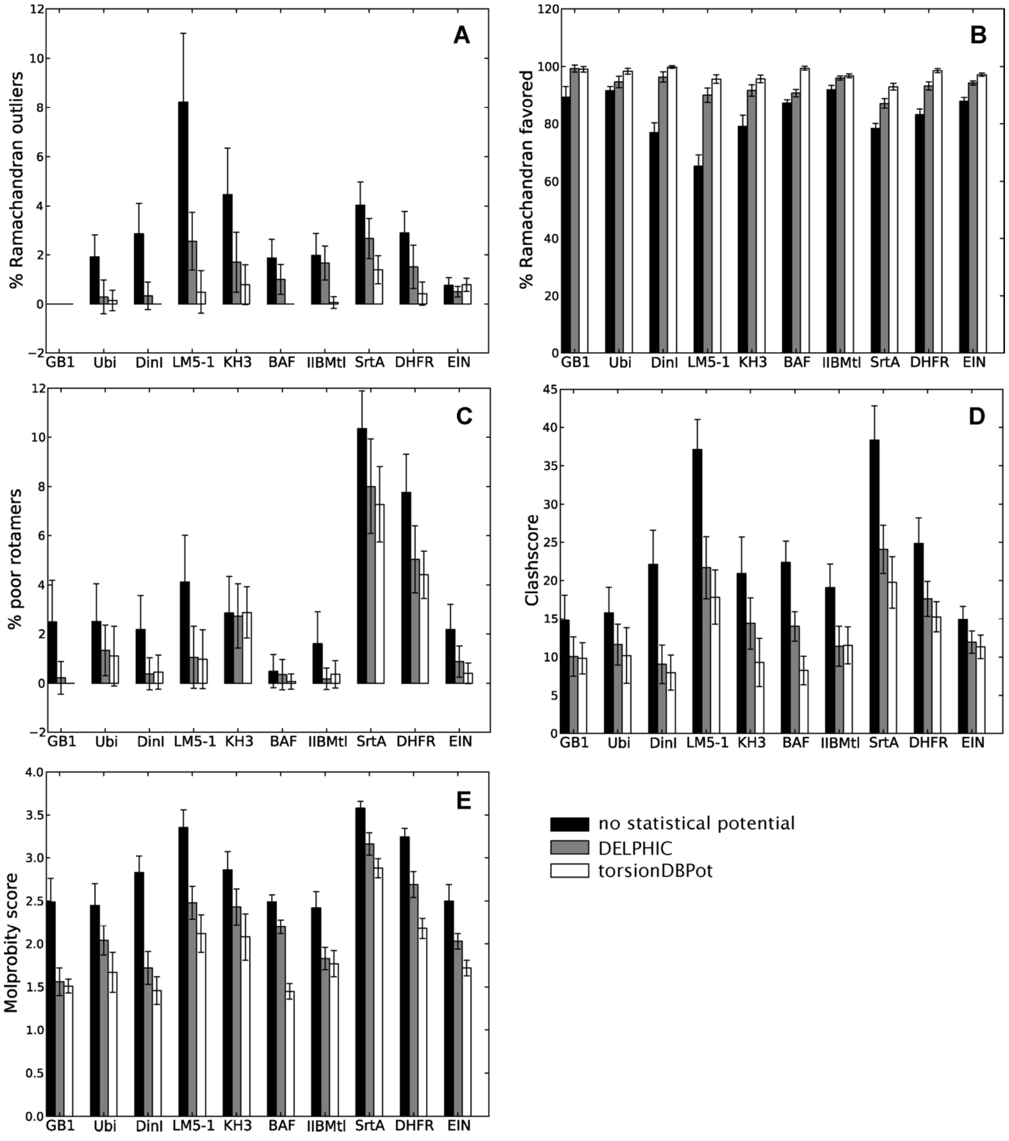
Figure 3. WHAT IF validation. Each barplot displays a WHAT IF validation statistic for structure ensembles of different proteins, with bars representing the mean ± standard deviation computed from 20 structures. Structure calculations without any statistical torsion angle potential (black), with the DELPHIC potential[12] (gray), and with the new torsionDBPot potential (white) are included. Abbreviated protein names are used; for full

names see Table II. Each statistic is a score: the larger the better. Packing quality (panel C) refers to the 2nd generation packing quality.

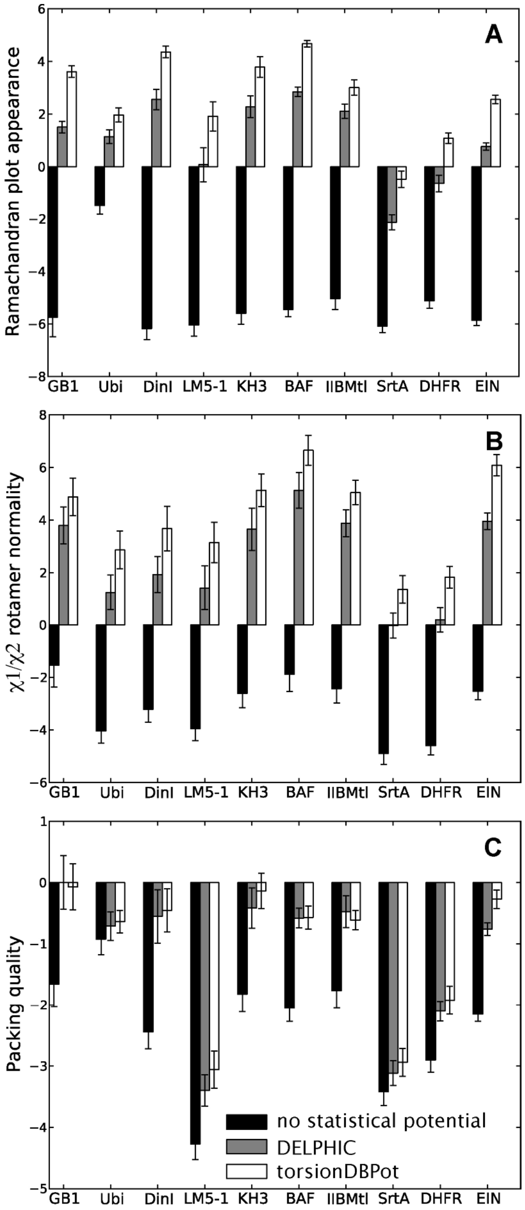
Figure 4. Fit to experimental data. Each barplot displays a figure of merit for the fit to a given experimental NMR observable of structure ensembles of different proteins, with bars representing the mean ± standard deviation computed from 20 structures (note that error bars associated with very small standard deviations may seem missing). Structure calculations without any statistical torsion angle potential (black), with the DELPHIC potential[12] (gray), and with the new torsionDBPot potential (white) are included. Abbreviated protein names are used; for full names see Table II. Large torsion angle RMS deviations for ubiquitin (panel B, asterisc) stem from unrealistically narrow bounds in the publicly released restraints (PDB ID: 1D3Z). Each RDC R-factor (panel C) is an unweighted average over different alignment media and nuclei pairs (when applicable).



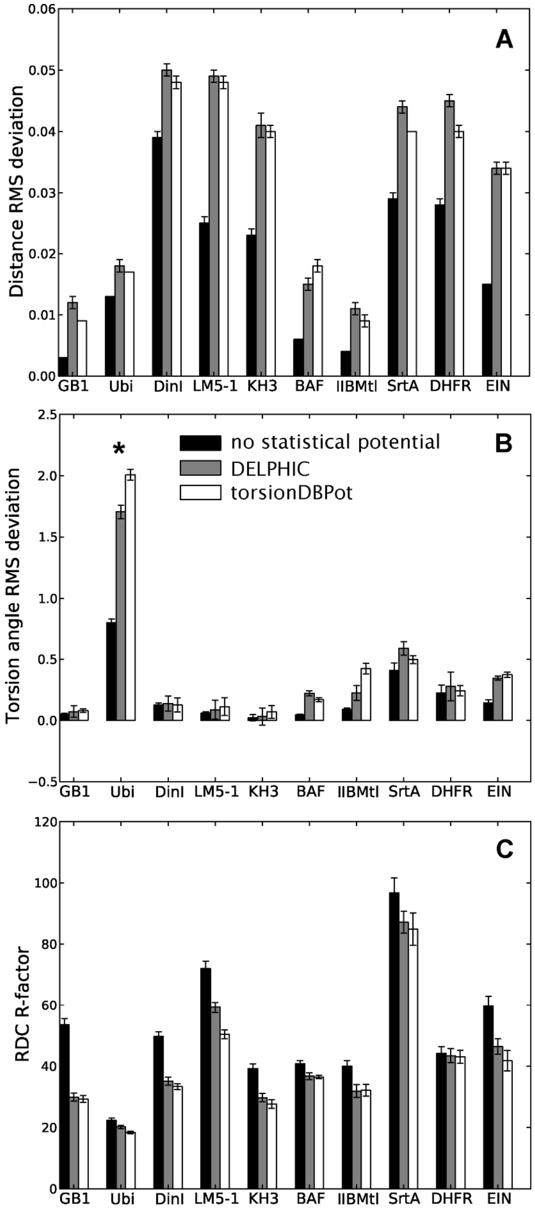
412x352mm (300 x 300 DPI)



162x183mm (300 x 300 DPI)



162x375mm (300 x 300 DPI)



162x370mm (300 x 300 DPI)

Smooth statistical torsion angle potential derived from a large conformational database via adaptive kernel density estimation improves the quality of NMR protein structures

Guillermo A. Bermejo, G. Marius Clore, and Charles D. Schwieters

Supporting Information

Table SI. Probability Density Expressions Extracted from the Torsion Angle Database

Residue Type	**Residues in Database** [a]	**Probability Density Function** [b]	**Statistical Approximation** [c]	**Energy Term Applied To:** [d]
Gly	All glycines (93,113)	$p(\phi,\psi)$	None	All glycines
cis-Pro	All cis-prolines (2,343)	$p(\phi,\psi,\chi_2)$	χ_2 determines other ring torsion angles	All cis-prolines
trans-Pro	All trans-prolines (46,686)	$p(\phi,\psi,\chi_2)$	χ_2 determines other ring torsion angles	All trans-prolines
$\text{prePro}_{\text{Ala}}$	All non-glycine, non-proline, pre-proline residues (39,093)	$p(\phi,\psi)$	None	pre-proline alanines
$\text{prePro}_{\text{Thr}}$	All non-glycine, non-proline, non-alanine, pre-proline residues (35,471)	$p(\phi,\psi,\chi_1)$	None	pre-proline threonines
$\text{prePro}_{\text{Val}}$	All non-glycine, non-proline, non-alanine, pre-proline residues (35,471)	$p(\phi,\psi,\chi_1)$	None	pre-proline valines
$\text{prePro}_{\text{Ser}}$	All non-glycine, non-proline, non-alanine, pre-proline residues (35,471)	$p(\phi,\psi,\chi_1)$	None	pre-proline serines
$\text{prePro}_{\text{Cys}}$	All non-glycine, non-proline, non-alanine, pre-proline residues (35,471)	$p(\phi,\psi,\chi_1)$	None	pre-proline cysteines
$\text{prePro}_{\text{Asp}}$	All non-glycine, non-proline, non-alanine, pre-proline residues (35,471) All aspartates (57,985)	$p(\phi,\psi,\chi_1)$ × $p(\chi_2)$	$\chi_2 \perp \phi,\psi,\chi_1$	pre-proline aspartates
$\text{prePro}_{\text{Asn}}$	All non-glycine, non-proline, non-alanine, pre-proline residues (35,471) All asparagines (44,158)	$p(\phi,\psi,\chi_1)$ × $p(\chi_2)$	$\chi_2 \perp \phi,\psi,\chi_1$	pre-proline asparagines
$\text{prePro}_{\text{Ile}}$	All non-glycine, non-proline, non-alanine, pre-proline residues (35,471) All isoleucines (59,385)	$p(\phi,\psi,\chi_1)$ × $p(\chi_2)$	$\chi_2 \perp \phi,\psi,\chi_1$	pre-proline isoleucines
$\text{prePro}_{\text{Leu}}$	All non-glycine, non-proline, non-alanine, pre-proline residues (35,471) All leucines (95,250)	$p(\phi,\psi,\chi_1)$ × $p(\chi_2)$	$\chi_2 \perp \phi,\psi,\chi_1$	pre-proline leucines
$\text{prePro}_{\text{His}}$	All non-glycine, non-proline, non-alanine, pre-proline residues (35,471) All histidines (24,061)	$p(\phi,\psi,\chi_1)$ × $p(\chi_2)$	$\chi_2 \perp \phi,\psi,\chi_1$	pre-proline histidines

Table SI. (Continued.)

Residue Type	Residues in Database [a]	Probability Density Function [b]	Statistical Approximation [c]	Energy Term Applied To: [d]
$\mathrm{prePro}_{\mathrm{Trp}}$	All non-glycine, non-proline, non-alanine, pre-proline residues (35,471)	$p(\phi,\psi,\chi_1)$ ×	$\chi_2 \perp \phi,\psi,\chi_1$	pre-proline tryptophans
	All tryptophans (15,671)	$p(\chi_2)$		
$\mathrm{prePro}_{\mathrm{Tyr}}$	All non-glycine, non-proline, non-alanine, pre-proline residues (35,471)	$p(\phi,\psi,\chi_1)$ ×	$\chi_2 \perp \phi,\psi,\chi_1$	pre-proline tyrosines
	All tyrosines (38,484)	$p(\chi_2)$		
$\mathrm{prePro}_{\mathrm{Phe}}$	All non-glycine, non-proline, non-alanine, pre-proline residues (35,471)	$p(\phi,\psi,\chi_1)$ ×	$\chi_2 \perp \phi,\psi,\chi_1$	pre-proline phenylalanines
	All phenylalanines (45,572)	$p(\chi_2)$		
$\mathrm{prePro}_{\mathrm{Met}}$	All non-glycine, non-proline, non-alanine, pre-proline residues (35,471)	$p(\phi,\psi,\chi_1)$ ×	$\chi_2,\chi_3 \perp \phi,\psi,\chi_1$	pre-proline methionines
	All methionines (14,319)	$p(\chi_2,\chi_3)$		
$\mathrm{prePro}_{\mathrm{Glu}}$	All non-glycine, non-proline, non-alanine, pre-proline residues (35,471)	$p(\phi,\psi,\chi_1)$ ×	$\chi_2,\chi_3 \perp \phi,\psi,\chi_1$	pre-proline glutamates
	All glutamates (46,950)	$p(\chi_2,\chi_3)$		
$\mathrm{prePro}_{\mathrm{Gln}}$	All non-glycine, non-proline, non-alanine, pre-proline residues (35,471)	$p(\phi,\psi,\chi_1)$ ×	$\chi_2,\chi_3 \perp \phi,\psi,\chi_1$	pre-proline glutamines
	All glutamines (31,349)	$p(\chi_2,\chi_3)$		
$\mathrm{prePro}_{\mathrm{Arg}}$	All non-glycine, non-proline, non-alanine, pre-proline residues (35,471)	$p(\phi,\psi,\chi_1)$ ×	$\chi_2,\chi_3,\chi_4 \perp \phi,\psi,\chi_1$	pre-proline arginines
	All arginines (38,701)	$p(\chi_2,\chi_3,\chi_4)$		
$\mathrm{prePro}_{\mathrm{Lys}}$	All non-glycine, non-proline, non-alanine, pre-proline residues (35,471)	$p(\phi,\psi,\chi_1)$ ×	$\chi_2,\chi_3,\chi_4 \perp \phi,\psi,\chi_1$	pre-proline lysines
	All lysines (36,834)	$p(\chi_2,\chi_3,\chi_4)$		

Table SI. (Continued.)

Residue Type	Residues in Database [a]	Probability Density Function [b]	Statistical Approximation [c]	Energy Term Applied To: [d]
Ala	All non-pre-proline alanines (98,680)	$p(\phi,\psi)$	None	non-pre-proline alanines
Thr	All non-pre-proline threonines (58,953)	$p(\phi,\psi,\chi_1)$	None	non-pre-proline threonines
Val	All non-pre-proline valines (75,992)	$p(\phi,\psi,\chi_1)$	None	non-pre-proline valines
Ser	All non-pre-proline serines (58,156)	$p(\phi,\psi,\chi_1)$	None	non-pre-proline serines
Cys	All non-pre-proline cysteines (9,577)	$p(\phi,\psi,\chi_1)$	None	non-pre-proline cysteines
Asp	All non-pre-proline aspartates (54,955)	$\frac{p(\phi,\psi,\chi_1)p(\chi_1,\chi_2)}{p(\chi_1)}$	$\chi_2 \perp \phi,\psi$ given χ_1	non-pre-proline aspartates
Asn	All non-pre-proline asparagines (41,623)	$\frac{p(\phi,\psi,\chi_1)p(\chi_1,\chi_2)}{p(\chi_1)}$	$\chi_2 \perp \phi,\psi$ given χ_1	non-pre-proline asparagines
Ile	All non-pre-proline isoleucines (56,534)	$\frac{p(\phi,\psi,\chi_1)p(\chi_1,\chi_2)}{p(\chi_1)}$	$\chi_2 \perp \phi,\psi$ given χ_1	non-pre-proline isoleucines
Leu	All non-pre-proline leucines (90,123)	$\frac{p(\phi,\psi,\chi_1)p(\chi_1,\chi_2)}{p(\chi_1)}$	$\chi_2 \perp \phi,\psi$ given χ_1	non-pre-proline leucines
His	All non-pre-proline histidines (22,619)	$\frac{p(\phi,\psi,\chi_1)p(\chi_1,\chi_2)}{p(\chi_1)}$	$\chi_2 \perp \phi,\psi$ given χ_1	non-pre-proline histidines
Trp	All non-pre-proline tryptophans (15,110)	$\frac{p(\phi,\psi,\chi_1)p(\chi_1,\chi_2)}{p(\chi_1)}$	$\chi_2 \perp \phi,\psi$ given χ_1	non-pre-proline tryptophans
Tyr	All non-pre-proline tyrosines (36,671)	$\frac{p(\phi,\psi,\chi_1)p(\chi_1,\chi_2)}{p(\chi_1)}$	$\chi_2 \perp \phi,\psi$ given χ_1	non-pre-proline tyrosines
Phe	All non-pre-proline phenylalanines (43,564)	$\frac{p(\phi,\psi,\chi_1)p(\chi_1,\chi_2)}{p(\chi_1)}$	$\chi_2 \perp \phi,\psi$ given χ_1	non-pre-proline phenylalanines
Met	All non-pre-proline methionines (13,727)	$\frac{p(\phi,\psi,\chi_1)p(\chi_1,\chi_2,\chi_3)}{p(\chi_1)}$	$\chi_2 \perp \phi,\psi$ given χ_1 $\chi_3 \perp \phi,\psi$ given χ_1,χ_2	non-pre-proline methionines
Glu	All non-pre-proline glutamates (45,332)	$\frac{p(\phi,\psi,\chi_1)p(\chi_1,\chi_2,\chi_3)}{p(\chi_1)}$	$\chi_2 \perp \phi,\psi$ given χ_1 $\chi_3 \perp \phi,\psi$ given χ_1,χ_2	non-pre-proline glutamates
Gln	All non-pre-proline glutamines (30,031)	$\frac{p(\phi,\psi,\chi_1)p(\chi_1,\chi_2,\chi_3)}{p(\chi_1)}$	$\chi_2 \perp \phi,\psi$ given χ_1 $\chi_3 \perp \phi,\psi$ given χ_1,χ_2	non-pre-proline glutamines
Arg	All non-pre-proline arginines (37,112)	$\frac{p(\phi,\psi,\chi_1)p(\chi_1,\chi_2,\chi_3)p(\chi_2,\chi_3,\chi_4)}{p(\chi_1)p(\chi_2,\chi_3)}$	$\chi_2 \perp \phi,\psi$ given χ_1 $\chi_3 \perp \phi,\psi$ given χ_1,χ_2 $\chi_4 \perp \phi,\psi$ given χ_2,χ_3	non-pre-proline arginines
Lys	All non-pre-proline lysines (35,281)	$\frac{p(\phi,\psi,\chi_1)p(\chi_1,\chi_2,\chi_3)p(\chi_2,\chi_3,\chi_4)}{p(\chi_1)p(\chi_2,\chi_3)}$	$\chi_2 \perp \phi,\psi$ given χ_1 $\chi_3 \perp \phi,\psi$ given χ_1,χ_2 $\chi_4 \perp \phi,\psi$ given χ_2,χ_3	non-pre-proline lysines

[a] Residues in the torsion angle database used to estimate the probability density function (the total number of instances is indicated in parenthesis).

[b] Expression used to represent the full joint probability density function. For simplicity, conditionals are omitted (for residue type Gly, $p(\phi, \psi)$ stands for $p(\phi, \psi|\text{Gly})$, etc.). For certain prePro residue subtypes (e.g., prePro$_{\text{Asp}}$) the different probability density components arise from different residue populations in the database.

[c] Approximation used to break down the full probability density function into components of lower dimensionality ($\perp$ means "independent of").

[d] Residues to which the associated statistical potential is applied during structure calculations.